nature genetics

A genome-wide scalable SNP genotyping assay using microarray technology

Kevin L Gunderson[1], Frank J Steemers[1,4], Grace Lee[1,4], Leo G Mendoza[2] & Mark S Chee[3]

Oligonucleotide probe arrays have enabled massively parallel analysis of gene expression levels from a single cDNA sample. Application of microarray technology to analyzing genomic DNA has been stymied by the sequence complexity of the entire human genome. A robust, single base–resolution direct genomic assay would extend the reach of microarray technology. We developed an array-based whole-genome genotyping assay that does not require PCR and enables effectively unlimited multiplexing. The assay achieves a high signal-to-noise ratio by combining specific hybridization of picomolar concentrations of whole genome–amplified DNA to arrayed probes with allele-specific primer extension and signal amplification. As proof of principle, we genotyped several hundred previously characterized SNPs. The conversion rate, call rate and accuracy were comparable to those of high-performance PCR-based genotyping assays.

Array technology has enabled whole-genome gene-expression studies of tens of thousands of genes across a single sample. Bottlenecks in sample preparation have precluded the application of whole-genome arrays to genotype analysis. This is unfortunate because the number of genotyping assays required for comprehensive association studies far outstrips the capacity of current approaches. The human genome contains >10 million common SNPs. A smaller number of well-selected 'tagging' SNPs, which act as markers of more extensive patterns of common variation, or haplotypes, can be used to map most of the genetic variation between individuals[1–4]. Preliminary estimates indicate that ∼200,000–300,000 tagging SNPs will be required to map most of the variation in the genome, depending on the population to be studied[4–6]. Some parts of the genome may not be amenable to this approach and may require a much higher density of SNPs. Therefore, even the most conservative estimates of the number of SNP markers required for comprehensive genetic association studies indicate that genotyping of several hundred thousand SNPs per individual across hundreds to many thousands of samples would be required to map a causal variant by linkage disequilibrium. The development of a direct array-based SNP genotyping assay and single-tube sample preparation would make it possible to genotype hundreds of thousands of SNPs efficiently. This would eliminate the need for PCR amplification, which may require many separate reactions for whole-genome studies[7,8].

Although human genomic DNA (gDNA) has been used successfully in cDNA, BAC and oligonucleotide array comparative genomic hybridization experiments[9–11], specific hybridization of the entire human genome to a high-density oligonucleotide probe microarray providing single-base resolution suitable for genotyping has not been reported. Sequence complexity is a barrier, because genomes of less complex organisms have been successfully genotyped on arrays. Winzeler *et al.* fragmented and hybridized the 12-Mb yeast genome to 25-mer oligonucleotide probe arrays designed for gene-expression profiling and discovered and genotyped biallelic markers[12]. A similar array-based oligonucleotide hybridization approach was used by Borevitz *et al.* to characterize nucleotide variation in *Arabidopsis* strains (120-Mb genome)[13]. Recent experiments indicate that genotyping directly from human gDNA is possible. Oligonucleotide probes are capable of sequence-specific hybridization in the context of the entire human genome[14]. Storhoff *et al.* specifically detected a single human locus directly from gDNA (∼200 fM concentration) using gold nanosphere detection[15]. The potential of genotyping directly from gDNA without intermediate amplification has been demonstrated by collecting genotype information from a single human locus from gDNA captured on a microsphere support, albeit with low signal-to-noise ratio[16]. More recently, Chen *et al.* similarly demonstrated genotyping of a single SNP locus directly from gDNA[17]. Though encouraging, robust array-based genotyping with these approaches has not yet been accomplished.

Given this genomic complexity challenge, the most common strategy is to reduce complexity by amplifying portions of the genome by PCR. This can be done in a random or semirandom fashion, for example, by using restriction enzyme–based adapter ligation PCR[18]. This representational approach, used to detect changes in genomic copy number, has also been applied to genotyping of 10^4–10^5 or more SNPs from a single sample preparation[19,20]. Highly multiplexed PCR-based approaches have recently been developed to genotype specific loci of interest, enabling targeted genotyping of ∼10^3–10^4 or more SNPs from a single sample[21,22]. These targeted genotyping approaches

[1]Illumina, Inc., 9885 Towne Centre Dr., San Diego, California 92121, USA. [2]Ambion, Inc., 2130 Woodward, Austin, Texas 78744, USA. [3]Prognosys Biosciences, Inc., 4215 Sorrento Valley Blvd., Suite 105, San Diego, California 92121, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to K.L.G. (kgunderson@illumina.com).

Published online 17 April 2005; doi:10.1038/ng1547

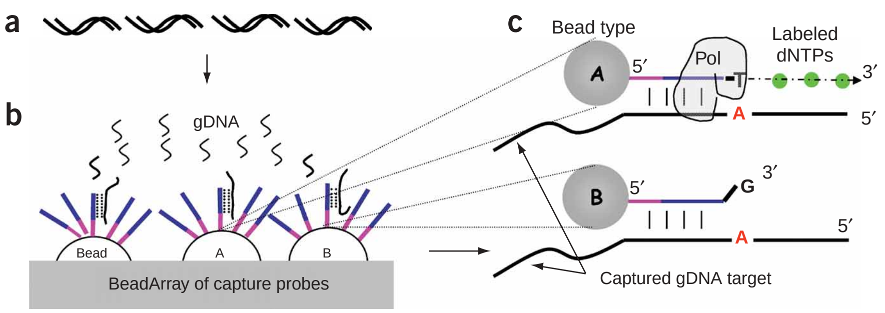

Figure 1 WGG on DNA arrays. (**a**) WGA to generate large amounts of amplified complex gDNA. (**b**) Hybridization of the WGA product to a specific and sensitive oligonucleotide probe array (50-mers). (**c**) An array-based ASPE reaction that scores the captured SNP targets by incorporating multiple biotin-labeled dNTP nucleotides into the appropriate allelic probe, followed by a sensitive detection and signal amplification step to read the incorporated labels. For a given SNP on a given strand, ASPE uses two different allele-specific bead types whose capture sequence is identical except for the 3′ terminal base. The base is chosen such that the probes on one bead type match allele A and those on the other bead type match allele B. Polymerase extension occurs preferentially from matched 3′ termini, enabling appropriate scoring of the SNP.

include a pre-PCR enzymatic SNP scoring step, addition of universal priming sites by ligation and subsequent universal primer PCR. Today, almost all high-throughput genotyping is done using one of these methods. The ability of these methods to scale to 10^5–10^6 genotypes per sample is uncertain.

Here we describe a new array-based whole-genome genotyping (WGG) assay with high signal-to-noise ratio, which allows for accurate and robust genotyping in the context of full genomic complexity. The WGG assay unlocks the potential of microarray technology to read hundreds of thousands to millions of SNPs per genome in a single array experiment. It mirrors an array-based gene-expression experiment in using single-tube sample preparation with virtually unlimited assay multiplex scalability, dependent only on the number of features on the array. Using this approach, it is now feasible to read a vast amount of genetic information directly from the genome using simple, easily automated procedures.

RESULTS

Assay design

The design of the assay is shown in **Figure 1**. It uses a combination of well-proven technologies: (i) whole-genome amplification (WGA) to generate sufficient DNA for hybridization; (ii) hybridization capture of the amplified genomic loci to a specific and sensitive oligonucleotide probe array; (iii) array-based enzymatic SNP scoring assay; and (iv) sensitive signal amplification for readout.

It is crucial to hybridize a sufficiently high concentration (partial concentration of 2–3 pM) of WGA gDNA to the array. We started with an input of ~100–200 ng of gDNA to generate 100–150 μg (1,000–1,500× amplification) of high-complexity amplified product[23].

We assessed the representation bias of the WGA reactions by comparing assay intensities between unamplified and amplified gDNA across 480 nonpolymorphic loci (**Fig. 2**). More than 95% of the loci had signal intensities within threefold of that of the unamplified sample, well within the dynamic range of our assay. Therefore, we estimate that >95% of the unique portion of the genome is accessible with our approach. Moreover, the amplification bias was highly reproducible from sample to sample ($R^2 = 0.98$), ensuring that a set of selected functional assays will result in reproducible amplification from experiment to experiment. Furthermore, reproducible amplification bias is important for future DNA copy and loss-of-heterozygosity applications.

After amplification, we hybridized WGA samples to arrays of 50-base capture probes (actual oligonucleotide probes are 75 bases long, of which 25 bases are used for decoding). We chose this length as a compromise to maximize signal intensity, minimize interlocus intensity variation and synthesize full-length probes. After hybridization, we scored SNPs using an allele-specific primer extension (ASPE) assay, chosen for its proven high discrimination and the economy of using the capture probes themselves to query the SNP. The ASPE approach also enables us to query all SNP categories with a single color on a single array. We accomplished this by using two bead types (A and B) per assay whose probe sequences differed only at the 3′ terminal base (opposite SNP site), creating allelic discrimination in the polymerase extension step. Probe sequences on a particular bead type will be extended and labeled (multiple incorporation events) only when hybridized to a perfectly matched allelic target. The genotype state of a given SNP locus (AA, AB or BB) is determined by the intensity ratio between the two corresponding bead types, $\theta = (2/\pi)*\arctan(B/A)$.

Assay optimization

For assay optimization, we designed ASPE assays for a set of 96 nonpolymorphic loci (a subset of the 480 loci) randomly selected from throughout the human genome. We used an artificial mismatch probe and an arbitrarily chosen 3′ mismatch base. Monitoring the assay discrimination (perfect match-to-mismatch ratio) of the

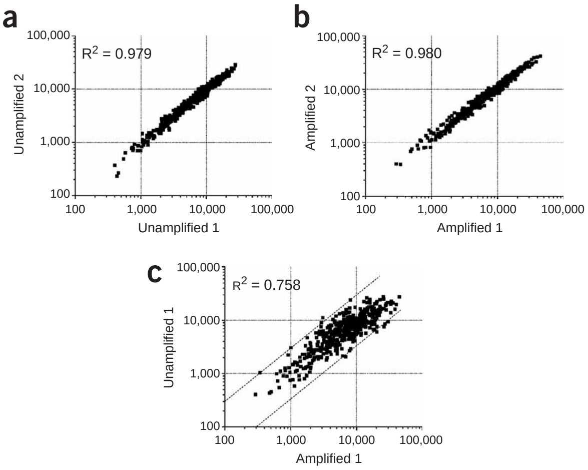

Figure 2 WGA representation. Assay intensities from the 480 nonpolymorphic controls were used to assess representation bias of amplified DNA. Scatter plots of (**a**) unamplified gDNA versus unamplified gDNA, (**b**) amplified gDNA versus amplified gDNA and (**c**) unamplified gDNA versus amplified gDNA. The dashed lines denote threefold changes. The results indicate that >95% of the loci in the amplified product are within threefold of the copy number of the unamplified sample. Furthermore, the scatter plot of amplified gDNA versus amplified gDNA, **b**, indicates that the amplification bias is consistent between two different amplified samples.

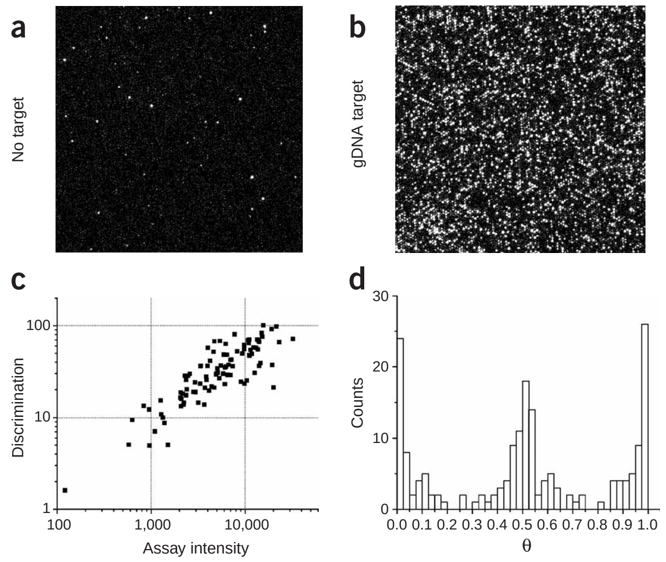

Figure 3 WGG on Sentrix array matrix using the WGG feasibility array. (**a**) The WGG assay was carried out on an array that went through hybridization without any DNA target. This 'no target control' assesses the self-extension of the probes on the array. (Most of the bright beads are detection controls included on every array to validate the signal amplification.) (**b**) The WGG assay was carried out on an array hybridized with amplified gDNA. (**c**) A scatter plot of discrimination ratio (match/mismatch) versus assay intensity across the 96 nonpolymorphic ASPE controls. The discrimination improves with assay intensity and has a median of ~30. (**d**) Histogram of the θ values for 186 SNP assays from a single sample. The histogram has three peaks corresponding to the three possible genotype categories.

nonpolymorphic controls in a Sentrix array matrix format (96 arrays) enabled us to optimize the WGA, hybridization and primer extension reactions[24]. Additional assays monitored locus representation (480 correctly matched nonpolymorphic sequences) and genotyping quality (186 ASPE linkage panel SNP assays).

Analysis of the nonpolymorphic ASPE controls showed a high discrimination ratio (median >30:1 or θ > 0.97) and minimal self-extension (**Fig. 3**). This level of discrimination also correlated with good histogram cluster separation in the θ values across the 186 linkage panel SNPs (**Fig. 3c**). We used this set to genotype 32 DNA samples in triplicate (except two samples genotyped in duplicate). Of the 186 assays, 176 (94.6%) were successful on the basis of cluster separation score (CSS) and visual inspection. The genotyping data on the 176 successful assays had a call rate of 99.7% (16,501 of 16,544), reproducibility of 99.96% (16,362 of 16,368) and concordance of 99.97% (7,918 of 7,920, genotyping concordance data available on only 16 of 32 samples). These results indicate that genotyping was highly accurate and reproducible, with a very low dropout rate, for the complexity (~3 × 10^9 bp) of the human genome.

Validation on HapMap quality control SNPs

We arbitrarily chose the 186 SNP assays from a set of 2,500 linkage panel SNPs empirically selected to perform well in the GoldenGate assay. Therefore, we tested a second set of SNPs that had not undergone empirical filtering. The HapMap project used a set of 1,500 randomly chosen SNPs for quality control purposes (The International HapMap Consortium, unpublished data). We designed assays for 1,485 of the 1,500 SNPs on both strands (excluding 15 insertion-deletion polymorphism SNPs for design reasons). We chose the best strand, as determined by the CSS, for each of the 1,485 assays and used it to genotype 95 samples (30 triplicate samples and 5 duplicate DNAs) from the CEPH plate of the HapMap project (**Fig. 4**). We assessed assay quality by measuring reproducibility in duplicate samples, heritability in parent-child trios and concordance of genotype calls with the GoldenGate assay. As a result of the random selection process, the 1,500-SNP panel contained a large number of ill-defined SNPs (multiple-hit loci, poor sequence, etc.) and nonpolymorphic loci. To avoid systematic errors due to confusion between a nonpolymorphic locus and a multiple-hit locus (resulting from cross-hybridization to multiple sequences in the genome), we defined successful assays as those in which both SNP alleles were detected (*i.e.*, the minor and the major allele). We currently use this approach in product development. Using this metric, we found that 819 of 1,500 SNPs had a minor allele. For comparison, the GoldenGate assay found that 864 of 1500 SNPs had a minor allele. This indicates that the WGG assay successfully develops assays at ~95% the rate of the GoldenGate platform (819 of 864). Of these 819 'successful' loci, genotype calls across the 95 DNA samples (also used in training) generated a call rate of 99.7%, a reproducibility of 99.99%, an inheritance of 99.97% and a concordance of 99.9% (**Table 1**). Notably, the accuracy of the WGG assay is similar to that of the GoldenGate assay, despite the fact that the complexity of a WGG sample is 50,000 times higher when hybridized to the array.

DISCUSSION

Genotyping on probe arrays is limited by sample preparation and assay design constraints and not by array technology. The inherent mismatch between assay scalability (low multiplex) and array scalability (high multiplex) led us to develop a WGG approach.

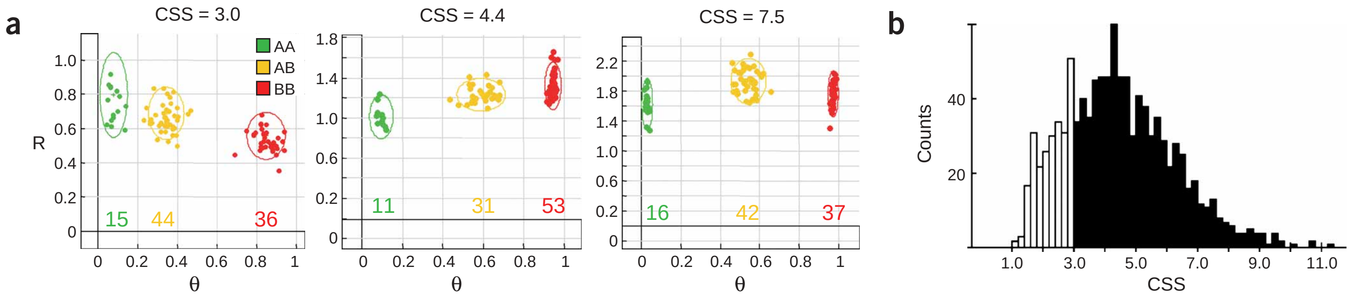

Figure 4 Genotyping of HapMap quality control SNPs. (**a**) Representative genotype plots illustrating different quality CSSs. This metric measures the statistical separation between genotype clusters and is used as a gauge for a successful assay. Loci with CSSs >3.0 are generally successful and provide reliable genotyping calls. (**b**) Histogram of CSSs for the 1,485 HapMap quality control loci. The median CSS across 1,485 assays was 4.4, and >76% of the assays had CSSs >3.0.

Table 1 Genotyping quality

Parameter	Value	Counts
Assay success*	95%	819 of 864
Call rate	99.5%	68,807 of 68,970
Reproducibility	99.99%	8,189 of 8,190
Heritability	99.97%	83,766 of 83,790
Concordance	99.9%	137,456 of 137,614

*Assay success rate computed relative to the GoldenGate assay.

Inherent in our assay design is the capacity for unlimited multiplexing, as each assay includes the full complexity of the human genome. Therefore, assay quality will be the same whether the assay queries a single SNP or one million SNPs. We carried out a proof-of-principle experiment on several hundred SNPs to demonstrate the concept. As a result of this new assay, the number of bases that can be analyzed in the unique portion of human gDNA is limited only by the number of probes in the array.

The WGG assay is modular in design. Each of the four key steps (WGA, array hybridization, array-based genotyping and signal amplification) can be replaced with alternatives. For instance, our current WGA procedure works well on relatively intact DNA (> 1–2 kb) but may not work well with highly degraded DNA (*e.g.*, from formalin-fixed paraffin-embedded samples). An alternative WGA method tolerant of sample degradation could be used[25]. Similarly, the enzymatic SNP scoring step can be replaced by alternatives such as single-base extension, ligation and allelic oligonucleotide hybridization. Finally, the signal amplification step can be replaced with alternative signal amplification schemes with enhanced sensitivity[26,27]. This modular design allows improvements to be quickly and easily incorporated into the assay.

We achieved a high signal-to-noise ratio by hybridizing relatively high concentrations (~2–3 pM) of WGA gDNA and by minimizing 'background' signal generated from interprobe and intraprobe self-extension. Self-extension was greatly suppressed by the addition of single-stranded binding protein to the extension reaction (data not shown). We achieved high specificity, in the presence of the entire genome, by combining stringent hybridization with on-array enzymatic discrimination. Enzymatic-based genotyping assays such as primer extension and ligation have previously been used directly on array surfaces to achieve high levels of discrimination in genotyping captured PCR products[28–32].

The BeadArray technology also contributes to the high quality of the WGG assay. First, each bead type is present in ~30 copies on average, and so each measurement is repeated multiple times, enabling removal of outliers and increased precision[24]. Second, all beads of a particular type are made at the same time, ensuring consistency. This eliminates an important source of variability between the two allele-specific probes, because ASPE-based genotyping uses a ratiometric comparison of the bead type intensities. Finally, we used 5′ immobilization of oligonucleotide probes synthesized in the 3′–5′ direction, yielding intact 3′ termini suitable for primer extension–based assays. Therefore, the exceptional performance of the WGG assay derives from a combination of design elements involving the assay as well as the readout platform.

In comparison to other high-throughput genotyping technologies, WGG combines virtually unlimited multiplexing with freedom of SNP selection. This will be particularly important in assaying custom sets of SNPs, including haplotype-tagging SNPs[33], coding SNPs and other high-value SNPs. Furthermore, the ability to choose SNPs of interest increases the power of association studies relative to random SNPs, particularly with respect to maximizing linkage disequilibrium with markers in feature-rich regions, such as genes and conserved regions[34]. This contrasts with the PCR-based complexity reduction approach, which, though capable of genotyping tens of thousands to hundreds of thousands of SNPs from a single sample or array, suffers from a limited ability to choose SNPs because it requires the SNP to be present in the amplified representation[19,20,35].

Other recently introduced genotyping technologies, such as the GoldenGate assay from Illumina and the Molecular Inversion Probe assay from Parallele, enable researchers to choose almost any SNP of interest and to multiplex to relatively large numbers (1,000–10,000)[21,22]. The ability to multiplex to even higher levels will, at some point, be limited by assay complexity and the amount of material generated for each individual locus in the final PCR reaction. These assays either do not allow genotyping of targeted SNPs or do not scale intrinsically like the WGG assay.

As a next step to realizing the full potential of the WGG assay, we are developing a single array to genotype > 100,000 loci from a one-tube reaction. This is made possible by the feature density of our arrays of > 288,000 bead types per slide. The BeadChip platform can be scaled to even higher densities. For instance, decreasing bead diameter from the current 3 μm to 1.5 μm (and reducing center-to-center spacing accordingly) enables use of > 1 million bead types per BeadChip.

In addition to standard SNP genotyping for association studies, these high-resolution genotyping arrays are applicable to genomic profiling (DNA copy and loss-of-heterozygosity measurements) and allelic expression measurements (cDNA genotyping). Partly as a result of this new assay, we anticipate a future in which massively parallel direct genomic analysis is used routinely in applications ranging from the exploration of genome function to molecular medicine.

METHODS

ASPE Sentrix array design. We obtained all the array data using Illumina's Sentrix BeadArray matrix. Oligonucleotide probes on the beads were 75 bases in length; 25 bases at the 5′ end were used for decoding[24] and the remaining 50 bases were locus-specific. We immobilized the oligonucleotides on activated beads using a 5′ amino group[24]. The WGG feasibility array contained probes for 186 SNP assays (186 probe pairs, allele A and allele B), 384 nonpolymorphic assays (match only), 96 nonpolymorphic discrimination assays (96 probe pairs, match and mismatch), 96 yeast loci (96 probe pairs, match and mismatch) and 7 control bead types (0, 1, 10, 100, 1,000, 10^4 and 10^5 labels per bead). We chose the set of 186 SNP assays randomly from a linkage panel of 2,500 autosomal SNP using an arbitrary strand. The HapMap study used four Sentrix array types (~750 assays per array) with assays designed to both strands. We chose the best strand, as determined by the CSS, for analysis of the HapMap quality control SNPs.

WGA sample preparation. We subjected DNA samples to WGA using commercially available reagents (Illumina MP1 and AMM). We carried out 100-μl WGA reactions yielding ~150 μg of product from 100 ng of input gDNA, indicating that amplification was greater than 1,000-fold. We obtained similar results using Repli-g kits (Qiagen) in accordance with the manufacturer's instructions using 100 ng of input gDNA. We carried out fragmentation by adding 9.5 μl of calcium chloride (5 mM), 30.5 μl of water and 0.025 U DNaseI (Invitrogen) to 50 μl of Repli-g product. We incubated the reaction for 15 min at 37 °C. We inactivated DNaseI by adding 10 μl of 0.1% SDS and heating it to 95 °C for 10 min. The average yield was ~50 μg per reaction, and the average product size after digestion, estimated by gel analysis, was ~100–200 bases (data not shown). We pooled two reactions to generate sufficient material for hybridization. After WGA, we purified the amplified products from nucleotides and primers by ethanol precipitation or by Montage ultrafiltration plates (Millipore) and quantified them by a picoGreen assay (Molecular Probes) or by measuring absorbance at 260 nm. We resuspended the precipitated or concentrated DNA at ~5–6 μg μl^{-1} in 1× hybridization buffer (1 M NaCl, 100 mM potassium-phosphate buffer (pH 7.5) and 0.1% Tween 20) supplemented with 20% formamide. We used sonicated human placental gDNA (Sigma) as unamplified DNA. In these studies, all gDNA samples were from high-quality DNA obtained from the Coriell Institute. The analysis of highly degraded DNA has not yet been undertaken.

Array hybridization. We denatured the resuspended WGA product at 95 °C for 5 min and then exposed it to the Sentrix array matrix, which we mated to a microtiter plate, submerging the fiber bundles in 15 μl of hybridization sample. We incubated the entire assembly for 14–18 h at 48 °C with shaking. After hybridization, we washed arrays in 1× hybridization buffer and 20% formamide at 48 °C for 5 min.

ASPE-based detection. Before carrying out the array-based primer extension reaction, we washed Sentrix array matrices for 1 min with wash buffer (33.3 mM NaCl, 3.3 mM potassium phosphate and 0.1% Tween-20, pH 7.6) and then incubated them for 15 min in 50 μl of ASPE reaction buffer (Illumina EMM, containing polymerase, a mix of biotin-labeled and unlabeled nucleotides, single-stranded binding protein, bovine serum albumin and appropriate buffers and salts) at 37 °C. After the reaction, we immediately stripped the arrays in freshly prepared 0.1 N NaOH for 2 min and then washed and neutralized them twice in 1× hybridization buffer for 30 s. We detected the biotin-labeled nucleotides incorporated during primer extension using a sandwich assay similar to that previously described[36]. We blocked the arrays at room temperature for 10 min in 1 mg ml^{-1} bovine serum albumin in 1× hybridization buffer and then washed them for 1 min in 1× hybridization buffer. We then stained the arrays with streptavidin-phycoerythrin solution (1× hybridization buffer, 3 μg ml^{-1} streptavidin-phycoerythrin (Molecular Probes) and 1 mg ml^{-1} bovine serum albumin) for 10 min at room temperature, washed the arrays with 1× hybridization buffer for 1 min and then counter-stained them with an antibody reagent (10 μg ml^{-1} biotinylated antibody to streptavidin (Vector Labs) in 1× PBST (137 mM NaCl, 2.7 mM KCl, 4.3 mM sodium phosphate, 1.4 mM potassium phosphate and 0.1% Tween-20) supplemented with 6 μg ml^{-1} goat normal serum) for 20 min. After counter-staining, we washed the arrays in 1× hybridization buffer and restained them with streptavidin-phycoerythrin solution for 10 min. We washed the arrays one final time in 1× hybridization buffer before imaging them in 1× hybridization buffer on a custom CCD-based BeadArray imaging system[24]. We extracted intensities using custom image analysis software.

Data analysis and genotype calls. We carried out genotype analysis using Illumina's GenCall software (version 1.0.14), which compared intensities between probes for allele A and allele B across a large number of samples to create archetypal clustering patterns. These patterns allowed the genotyping data to be assigned membership to clusters using a probabilistic model and allowed assignment of a corresponding GenCall score. For example, data points falling between two clusters were assigned a low probability score of being a member of either cluster and had a correspondingly low GenCall score.

We initially assessed cluster quality by evaluating the CSS, a measure of statistical separation between clusters. It is defined as

$$\mathrm{CSS} = \min\left(\frac{|\theta_{AB} - \theta_{AA}|}{|\sigma_{AB} + \sigma_{AA}|}, \frac{|\theta_{AB} - \theta_{BB}|}{|\sigma_{AB} + \sigma_{BB}|}\right).$$

Loci with cluster scores around the cutoff of 3.0 were visually evaluated and the training clusters refined by manual intervention. We chose the cutoff value of 3.0 for the CSS on the basis of our experience in minimizing strand concordance errors. We generally observe accurate genotyping when the CSS is >3.0. Loci with questionable clusters were scored as unsuccessful and excluded from further analysis.

For the analysis of concordance of HapMap quality control SNPs with the GoldenGate assay, we used 726 of 819 SNP assays. Of the 93 SNP assays that we excluded, 92 were not called by the GoldenGate assay, and one (rs3778464) had high systematic discordance (170 of 190 calls were discordant) with GoldenGate. We computed the heritability by tabulating inheritance errors across autosomal SNPs (798 of 819) on the set of trios (35) in the CEPH sample plate of the HapMap. Heritability was defined as (total genotypes called – inheritance errors)/total genotypes called.

URL. More information on the HapMap project is available at http://www.hapmap.org/.

ACKNOWLEDGMENTS

We thank our Illumina colleagues in manufacturing for providing oligos and arrays and L. Zhou for generating the nonpolymorphic control sequences used in assay development. This work was supported in part by a US National Institutes of Health National Cancer Institute grant to K.L.G.

AUTHORS' CONTRIBUTIONS

K.L.G. conceived the experimental approach, oversaw the design and analysis of experiments and, together with M.S.C., wrote the paper. F.J.S., G.L. and L.G.M. carried out the experiments, analyzed data and identified and implemented a number of technical improvements.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 18 October 2004; accepted 10 March 2005
Published online at http://www.nature.com/naturegenetics/

1. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
2. Johnson, G.C. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237 (2001).
3. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 229–232 (2001).
4. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
5. Judson, R., Salisbury, B., Schneider, J., Windemuth, A. & Stephens, J.C. How many SNPs does a genome–wide haplotype map require? *Pharmacogenomics* **3**, 379–391 (2002).
6. Stephens, J.C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
7. Kwok, P.Y. & Chen, X. Detection of single nucleotide polymorphisms. *Curr. Issues Mol. Biol.* **5**, 43–60 (2003).
8. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
9. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**, 207–211 (1998).
10. Pollack, J.R. *et al.* Genome–wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**, 41–46 (1999).
11. Carvalho, B., Ouwerkerk, E., Meijer, G.A. & Ylstra, B. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J. Clin. Pathol.* **57**, 644–646 (2004).
12. Winzeler, E.A. *et al.* Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197 (1998).
13. Borevitz, J.O. *et al.* Large-scale identification of single–feature polymorphisms in complex genomes. *Genome Res.* **13**, 513–523 (2003).
14. Wu, D.Y., Nozari, G., Schold, M., Conner, B.J. & Wallace, R.B. Direct analysis of single nucleotide variation in human DNA and RNA using in situ dot hybridization. *DNA* **8**, 135–142 (1989).
15. Storhoff, J.J. *et al.* Diagnostic detection systems based on gold nanoparticle probes. in *Biomedical Applications of Micro- and Nanoengineering* Proc. SPIE vol. 4937 (ed. Nicolau, D.V.) 1–7 (SPIE, Bellingham, Washington, 2002).
16. Rao, K.V. *et al.* Genotyping single nucleotide polymorphisms directly from genomic DNA by invasive cleavage reaction on microspheres. *Nucleic Acids Res.* **31**, e66 (2003).
17. Chen, Y., Shortreed, M.R., Peelen, D., Lu, M. & Smith, L.M. Surface amplification of invasive cleavage products. *J. Am. Chem. Soc.* **126**, 3016–3017 (2004).
18. Lucito, R. *et al.* Genetic analysis using genomic representations. *Proc. Natl. Acad. Sci. USA* **95**, 4487–4492 (1998).
19. Kennedy, G.C. *et al.* Large-scale genotyping of complex DNA. *Nat. Biotechnol.* **21**, 1233–1237 (2003).
20. Matsuzaki, H. *et al.* Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.* **14**, 414–425 (2004).
21. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).

22. Fan, J.B. *et al.* Highly parallel SNP genotyping. *Cold Spring Harbor Symposia on Quantitative Biology LXVIII*, 69–78 (CSHL, Woodbury, New York, 2003).
23. Dean, F.B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA* **99**, 5261–5266 (2002).
24. Gunderson, K.L. *et al.* Decoding randomly ordered DNA arrays. *Genome Res.* **14**, 870–877 (2004).
25. Wang, G. *et al.* DNA amplification method tolerant to sample degradation. *Genome Res.* **14**, 2357–2366 (2004).
26. Bobrow, M.N., Harris, T.D., Shaughnessy, K.J. & Litt, G.J. Catalyzed reporter deposition, a novel method of signal amplification. Application to immunoassays. *J. Immunol. Methods* **125**, 279–285 (1989).
27. Hacker, G.W. High performance Nanogold-silver in situ hybridisation. *Eur. J. Histochem.* **42**, 111–120 (1998).
28. Shumaker, J.M., Metspalu, A. & Caskey, C.T. Mutation detection by solid phase primer extension. *Hum. Mutat.* **7**, 346–354 (1996).
29. Gunderson, K.L. *et al.* Mutation detection by ligation to complete n-mer DNA arrays. *Genome Res.* **8**, 1142–1153 (1998).
30. Pastinen, T., Kurg, A., Metspalu, A., Peltonen, L. & Syvanen, A.C. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* **7**, 606–614 (1997).
31. Pastinen, T. *et al.* A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res.* **10**, 1031–1042 (2000).
32. Erdogan, F., Kirchner, R., Mann, W., Ropers, H.H. & Nuber, U.A. Detection of mitochondrial single nucleotide polymorphisms using a primer elongation reaction on oligonucleotide microarrays. *Nucleic Acids Res.* **29**, E36 (2001).
33. Consortium, T.I.H. The International HapMap Project. *Nature* **426**, 789–796 (2003).
34. Simpson, C.L. *et al.* MaGIC: a program to generate targeted marker sets for genome-wide association studies. *Biotechniques* **37**, 996–999 (2004).
35. Di, X. *et al.* Dynamic model-based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* advance online publication, 19 January 2005 (doi:10.1093/bioinformatics/bti275).
36. Pinkel, D., Straume, T. & Gray, J.W. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc. Natl. Acad. Sci. USA* **83**, 2934–2938 (1986).